Dependence of Body Mass Index on Some Dietary Habits: An Application of Classification and Regression Tree

*Magdalena Platikanova¹, Antoaneta Yordanova², Petya Hristova¹

1. Department of Hygiene, Epidemiology, Microbiology, Parasitology and Infectious Diseases, Medical Faculty, Trakia University, Stara Zagora 6000, Bulgaria
2. Medical College, Trakia University, Stara Zagora 6000, Bulgaria

*Corresponding Author: Email: mplatikanova@abv.bg

(Received 17 Jul 2021; accepted 21 Sep 2021)

Abstract
Background: The purpose of this study was to determine the influence of some eating habits on body mass index (BMI) using a regression model created via the classification and regression tree method (CART).
Methods: The study was conducted using a questionnaire specially developed for the study, evaluated for reliability and validity. In addition to demographics (age and sex), the questions concern the timing of the meals and the type of food consumed. The data contains records for 533 people (322 women and 211 men) aged 18 to 65 years. The survey was conducted in the period 2019-2021 in Stara Zagora, Bulgaria. Data were processed using descriptive statistics, and regression and classification data mining method CART.
Results: A CART model with a dependent variable BMI and predictors Sex, Age, Breakfast type, Breakfast time, Lunchtime, Lunch type, Dinner time, Dinner type have been created. The obtained model is statistically significant at a significance level of P<0.0001 and a coefficient of determination R² = 0.495. The normalized importance of the factors that affect the BMI is as follows: Sex (100%), Age (61.4%), Lunch type (26.0%), Lunchtime (18.8%), Dinner time (13.9%), and Breakfast type (13.2%). Women have a lower BMI than men. BMI increases with age.
Conclusion: The CART method allows to make a classification by the predictors used and gives opportunities for a more in-depth analysis of the reasons for the increase in BMI. The level of influence of diet and eating habits (type of food, time of consumption) on BMI was determined.

Keywords: Body mass index (BMI); Dietary habits; Classification and regression tree (CART method)

Introduction

Nutrition plays a key role in human health (1-3). It is the main energy source and can contribute to a good quality of life (4, 5). The anthropometric indicator- body mass index (BMI) is most often used to assess the nutritional and health status. BMI refers to people with low physical activity. BMI is directly related to eating habits. Consumption of plant-based foods and fish is associated with lower BMI (6-8). Meat and foods rich in carbohydrates lead to increased BMI (6,9). Skipping breakfast in the morning also affects BMI. According to scientists, skipping breakfast leads to increased BMI (10). Late dinner also increases BMI (11). People who dine late
skip breakfast more often and this leads to weight gain (12). According to WHO, BMI was developed as an indicator of health status (13). As BMI increases, the risk of developing diabetes, obesity, cardiovascular, endocrine, cancer, and other diseases increases (14-18). A study by WHO predicts that by 2025, obesity may increase in 44 countries (19).

It is important to study how factors such as age, sex, eating habits affect BMI. The purpose of this study was to determine the influence of some eating habits on BMI using a regression model created via the classification and regression tree method (CART).

Materials and Methods

Data collection, structuring, and processing
The survey was conducted in the period 2019-2021. A questionnaire was developed for the study by the team that prepared the article. The concept of internal coherence has been applied to assess the reliability and validity of the test. The evaluation of the reliability of the questionnaire was carried out according to the Split-half method and through statistical procedures in the SPSS software package. The following values of the evaluation coefficients were obtained: Spearman-Brown Coefficient=0.730; Guttman split-half Coefficient=0.719, Cronbach's Alpha=0.479. The lower values of these coefficients can be explained by the specific scales for reporting the response levels. The validity was assessed through an expert evaluation of a pilot questionnaire by 10 specialists. Nine of them gave a positive assessment of the content. The questionnaire successfully measured what it was intended for. We conclude that the content of the questionnaire measured the research object. The data collection was following the demographic and economic profile of Stara Zagora region, where there is a structurally determining enterprise for electricity generation, a medium-sized university, some public and private enterprises. The questionnaire was filled in by 533 people, living in Stara Zagora, Bulgaria. The sample is random and includes 322 women and 211 men, aged 18 to 65 years. The survey was conducted by an interviewer. The research complies with the requirements for voluntary participation and confidentiality. The Ethics Committee of Medical Faculty, Trakia University, Stara Zagora, Bulgaria approved the research. Informed consent for participation in the research was developed according to Protocol № 9/15.05.2019, item 4. The answers were coded and applied in the data processing as ordinal or nominal values. Several sets of questions were included to establish the relationship between health and eating habits. BMI was calculated by the definition of WHO using and Measuring height and weight formula $\text{BMI} = \frac{W}{h^2}$, where $W$ is a person's weight in kilograms and $h$ is the height of the person in meters.

The study focus is on the type of breakfast, lunch, and dinner and the time of their consumption. The various food choices reflect both traditional and vegetarian diets. They are in line with the region's eating habits and new eating trends.

The coding application is as follows: Pastry breakfast is based on dough and pastry crusts. Dairy products include fresh milk and yogurt. The Sandwich group includes a burger, a doner. A meat dish means a dish cooked with meat, which can be a barbecue with salad or french fries. It can be meat with legumes, meat with vegetables, meat with pasta. These combinations are typical of traditional Bulgarian cuisine. There are 4 levels for breakfast, which are set from early to late breakfast. Lunchtime is limited to 3 levels -12, 13, and 14, following the traditional time for lunch break for working people. Dinner time levels are four and also set from early to late meals.

Statistical analysis
Data processing was performed using descriptive statistics and the CART method implemented in the SPSS, ver. 25 (IBM Corp., Armonk, NY,
A non-parametric CART modeling algorithm was used to obtain classification and regression models (18). BMI was treated as a dependent variable in the regression model. The independent variables in the model are Sex, Age, Breakfast type, Breakfast time, Lunchtime, Lunch type, Dinner time, Dinner type. The CART method was preferable because it did not require assumptions about the normal distribution of variables (20). The use of the CART method is popular in many areas and for studies among children, adolescents, and adults at high risk of obesity (21). Overweight and obesity in adolescence can be predicted by BMI using the classification and regression tree (CART) approach (22). This method is also used to explore associations between diet quality and socioeconomic parameters in children and adolescents (23). Various factors influence BMI and obesity. Prognostic factors for predicting obesity are childhood obesity, consumption of soft drinks, meats, pastries, fried food, low consumption of seafood (24-27). A review of the application of data mining methods, one of which is CART, in BMI research (28). CART method could also be used as a research tool for the identification of risk groups among populations in public health research (29).

Results

Descriptive statistics

The levels of BMI and nutritional status were determined as follows: below 18.5 (underweight), 18.5-24.9 (normal weight), 25.0-29.9 (overweight), 30.0-34.9 (obesity class I), 35.0-39.9 (obesity class II), and over 40 (obesity class III) (30). The data showed that 49.99% of the respondents were people with normal weight, overweight (26.64%), obesity class I (14.07%), underweight (6%), obesity class II (3%), and obesity class III (0.38%).

Table 1: Description of the cases by age and sex

<table>
<thead>
<tr>
<th>Demographic variables</th>
<th>Category</th>
<th>Number of cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (yr)</td>
<td>To 20</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>21-30</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td>31-40</td>
<td>127</td>
</tr>
<tr>
<td></td>
<td>41-50</td>
<td>148</td>
</tr>
<tr>
<td></td>
<td>51-60</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>Over 60</td>
<td>26</td>
</tr>
<tr>
<td>Sex</td>
<td>women</td>
<td>322</td>
</tr>
<tr>
<td></td>
<td>men</td>
<td>211</td>
</tr>
</tbody>
</table>

The respondents were divided into six groups by age: up to 20 yr, 21-30 yr, 31-40 yr, 41-50 yr, 51-60 yr, and over 60 years (Table 1). The average BMI values were respectively: 23.62, 22.93, 24.54, 26.08, 26.81, and 27.14. The highest level of BMI was in the age group over 60 years. In Fig. 1 the levels of BMI by age and sex are shown in a boxplot diagram. BMI increases with age. The mean BMI values for women and men were respectively 23.08 and 27.88. By Chi-Square Tests, there was a statistically significant difference in BMI between men and women with a significance level *** P<0.0001.
Creating a model

Using the CART method, a regression model was created with a dependent variable BMI and 8 independent variables: Sex, Age, Breakfast type, Breakfast time, Lunchtime, Lunch type, Dinner time, and Dinner type, their levels are described above.

The model did not include the variables Dinner type and Breakfast time, due to insufficient importance as predictors.

The following specifications were selected for model settings: minimum cases in parent node 15, minimum cases in child node 10, cross-validation -10, maximum tree depth 5. The resulting tree had 27 nodes, 14 of which were terminal. In the initial stage, all 533 observations were located in the root Node 0 of the tree with an average estimated value for BMI 24.998. The first division was by sex. All women were classified in the left Node - 322 cases with an expected mean BMI of 23.11. In the right Node, all the respondents were men -211 cases for which the expected (mean) value was 27.88. The following separation procedure continued from Node 1 by variable Age based on the mathematical criteria. The data in the two subsidiary Nodes 3 and 4 were calculated and presented, as in the first stage of the algorithm. Repeating this procedure, a regression tree was obtained as is shown in Fig. 2.

The analysis of the results showed that the overall mean relevant to the men was overweight and for women - normal. This can be seen from the descriptive statistics.

A full description of the rules used for the terminal nodes of the tree is presented in Table 2.
Fig. 2: The regression tree of the model
Table 2: Descriptions of the rules and BMI predicted values for the terminal nodes of the model decision tree

<table>
<thead>
<tr>
<th>Terminal Node</th>
<th>Number of cases</th>
<th>Rules</th>
<th>Mean/Predicted value</th>
<th>BMI code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node 18</td>
<td>89</td>
<td>Sex={Man}, Age&gt;24.5, Breakfast type={Pastry, Dairy products, Vegetables, Others}, Dinner time&gt;7pm</td>
<td>29.21</td>
<td>Overweight</td>
</tr>
<tr>
<td>Node 25</td>
<td>57</td>
<td>Sex={Man}, Age&gt;24.5, Breakfast={Pastry, Dairy products, Vegetables, Others}, Dinner time&lt;=7pm, Lunch time &lt;=12pm</td>
<td>28.56</td>
<td>Overweight</td>
</tr>
<tr>
<td>Node 26</td>
<td>26</td>
<td>Sex={Man}, Age&gt;24.5, Breakfast={Pastry, Dairy products, Vegetables, Others}, Dinner time&gt;7pm, Lunch time &gt;12pm</td>
<td>26.35</td>
<td>Overweight</td>
</tr>
<tr>
<td>Node 9</td>
<td>15</td>
<td>Sex={Man}, Age&lt;=24.5, Lunch type={Meat dish}</td>
<td>26.30</td>
<td>Overweight</td>
</tr>
<tr>
<td>Node 24</td>
<td>20</td>
<td>Sex={Woman}, Age&lt;=44.5, Age&lt;=44.5, Age&lt;=40.5, Breakfast type={Pastry, Others}</td>
<td>25.45</td>
<td>Overweight</td>
</tr>
<tr>
<td>Node 4</td>
<td>126</td>
<td>Sex={Woman}, Age&gt;44.5</td>
<td>25.33</td>
<td>Overweight</td>
</tr>
<tr>
<td>Node 11</td>
<td>10</td>
<td>Sex={Man}, Age&gt;24.5, Breakfast={Pastry, Dairy products, Vegetables, Others}, Dinner time&gt;7pm</td>
<td>24.89</td>
<td>Normal</td>
</tr>
<tr>
<td>Node 10</td>
<td>14</td>
<td>Sex={Man}, Age&lt;=24.5, Lunch type={Vegetarian dish, Sandwich, Soup}</td>
<td>23.29</td>
<td>Normal</td>
</tr>
<tr>
<td>Node 23</td>
<td>21</td>
<td>Sex={Woman}, Age&lt;=44.5, Age&lt;=44.5, Age&lt;=40.5, Breakfast type={Dairy products, Fruits}</td>
<td>22.76</td>
<td>Normal</td>
</tr>
<tr>
<td>Node 19</td>
<td>21</td>
<td>Sex={Woman}, Age&lt;=44.5, Age&lt;=36.5, Lunch type={Desert, Sandwich}, Dinner time&lt;=7pm, Sex={Woman}, Age&lt;=44.5, Age&lt;=36.5, Lunch type={Meat dish Vegetarian dish, Soup}, Breakfast type={Pastry, Vegetables, Others},</td>
<td>21.68</td>
<td>Normal</td>
</tr>
<tr>
<td>Node 22</td>
<td>44</td>
<td>Sex={Woman}, Age&lt;=44.5, Age&lt;=36.5, Lunch type={Meat dish Vegetarian dish, Soup}, Breakfast type={Pastry, Vegetables, Others},</td>
<td>21.64</td>
<td>Normal</td>
</tr>
<tr>
<td>Node 16</td>
<td>26</td>
<td>Sex={Woman}, Age&lt;=44.5, Age&gt;=36.5, Age&gt;=40.5</td>
<td>21.39</td>
<td>Normal</td>
</tr>
<tr>
<td>Node 21</td>
<td>44</td>
<td>Sex={Woman}, Age&lt;=44.5, Age&gt;=36.5, Age&gt;=36.5, Lunch type={Meat dish, Vegetarian dish, Soup}, Breakfast type={Dairy products, Fruits},</td>
<td>20.87</td>
<td>Normal</td>
</tr>
<tr>
<td>Node 20</td>
<td>20</td>
<td>Sex={Woman}, Age&lt;=44.5, Age&lt;=36.5, Lunch type={Desert, Sandwich}, Dinner time&gt;7pm</td>
<td>18.99</td>
<td>Normal</td>
</tr>
</tbody>
</table>

Table 2 presents the rules classifying cases for each terminal Node and the mean BMI in this node, which was the estimated value. For each terminal Node, the characteristics are described of the subjects in terms of age, sex, type of breakfast, and lunch and dinner time. These variables influence BMI. Many cases were in Node 4, which are women aged>44.5. These were the cases with the highest mean BMI in women. In node 20, BMI=18.99, although the lowest but it was normal. This node included young women under the age of 36.5, who prefer fresh food and sandwiches for lunch but eat after 7 pm. This group had a normal BMI. Lunch dietary habits separated the second group of women in another group aged under 36.5 with a higher BMI value=20.87. Women of the older age group 40.5 - 44.5 yr, also had normal BMI values=21.39 (Node 16). Node 4 shows that women over 44.5 yr had a mean value BMI = 25.33, which was in the lower limit of overweight. The highest BMI-25.45 was for these aged 36.5-40.5. They consumed mainly pastries. At the same age, women...
who had fruits and dairy had BMI=22.76. Age had the strongest influence on women. It was included in the regression tree three times. Eating pastry for breakfast leads to a higher BMI. 36.8% of women are under the age of 44.5. Their BMI was normal. For women over 44.5 BMI was high. For men, there were 6 end nodes and four of them (9, 25, 26, and 18) had a BMI excess, respectively 26.30, 26.35, 28.56, 29.21. In the groups of nodes 25, 26, and 28, men are over 24.5 yr old and eat mainly pastry and dairy products for breakfast, lunch, or dinner. These men had dinner very late or very early. Late dinner was the difference between nodes 25 and 26, as well as lunchtime, which led to an increase in BMI. The other two terminal nodes for men had normal values of BMI<25, Node 10 23.29 and Node 11-24.89. These respondents were under the age of 24.5 and had lighter meals, breakfast with fruit. The bottom line was that eating time was more important for men than women.

According to the normalized importance, the factors that affect the BMI were arranged as follows: Sex (100%), Age (61.4%), Lunch type (26.0%), Lunchtime (18.8%), Dinner time (13.9%), and Breakfast type (13.2%).

The created regression model is statistically significant at a significance level of $P<0.0001$. Diagnosis of model residuals shows that they have close to the normal distribution. Figure 3 shows a comparison of the values of the dependent variable BMI with those predicted by the CART model. A coefficient of determination $R^2 =0.495$ was obtained, which means that the model explains 50% of the change in BMI with the independent variables used.

**Fig. 3:** Comparison of the values of the dependent variable with the values predicted by the CART model with a 5% confidence interval

**Analysis of variables**

The percentage of fruits and vegetables consumed was low. Pastries predominate for all ages. More than 50% of women under the age of 20 consume pastry. Less than 10% of respondents...
had fruit for breakfast and less than 1% vegetables.
In the age group, 21-30 under 30% eat pastry for breakfast, followed by dairy products (about 20% for women and about 5% for men). About 5% of both sexes eat fruits for breakfast.
Overall, 20% of both sexes in the age group 31-40 consume pastries. In this age group, women consume dairy products more often - under 20%. Less than 10% of men eat dairy products for breakfast. About 10% of women and less than 5% of men eat fruits for breakfast. About 1% of women and only exceptional cases of men eat vegetables for breakfast.
Over 20% among both sexes aged 41-50 consume pastries, followed by dairy products - under 10%. In this age group, women more often consume fruits for breakfast - over 10%.
Respondents aged 51-60 also most often consumed pastries - more than 30% of women and about 25% of men. About 10% of women and less than 5% of men consume dairy products. Less than 10% of both sexes eat fruit for breakfast and vegetables - only exceptional cases.
Over 30% of women and about 20% of men aged over 60 consumed often pastry. Over 10% of women and over 20% of men consumed dairy products. About 10% of women and about 5% of men over 60 eat fruit for breakfast. Vegetables were rarely chosen as breakfast in this age group. (Fig. 4).
Figure 5 shows that men predominantly eat meat, and women frequently consume soups for lunch. This result was observed among the youngest women over 20 (over 30%). Over 20% of women aged 21-30 eat sandwiches for lunch. 15% of women aged 41-50 often consume vegetarian food. Men rarely consume vegetarian food – about 5% of these aged over 60.
Late dinner after 9 pm is typical for men of all ages. For 20% of elder women is typical earlier dinner between 6 pm to 7 pm. Men in the age groups 21-30 yr (over 20%), 31-40 (over 10%), and 41-50 (under 20%) dine late (after 9 pm.).
Discussion

The results of the study show that men have a higher BMI than women. Similar data are reported from other studies by scientists, who report a higher BMI among men than women (31,32). Opinions of scientists have established for its population a higher BMI and a greater incidence of overweight and obesity among women (33,34). Studies show that with age the BMI increases, which supports our results (35,36). Men of the same age is overweight or obese more often compared to women. With age, BMI changes inevitably (37,38).

An important factor related to BMI was dinner time. BMI is characteristic mostly of people who dine more later (39,40). A US study found that people aged 51-69 had higher BMI, supports our results (41). The results from our study are confirmed by English authors. They found that the intake of white bread increases BMI and the risk of obesity (42,43). Fruit consumption reduces BMI and the risk of overweight and obesity (44,45). Vegetarian food is preferred by women (46,47).

Conclusion

The main indicator for evaluation, BMI, is especially important. It depends on some factors such as sex, age, diet, and dietary habits. Women had a lower body mass index than men, which increases with age. In terms of some eating habits, BMI was affected by both the type of food and the time of consumption.

The relationship between BMI and other factors should continue to be studied in depth to improve eating behavior on time and reduce the risk of disease.

Journalism Ethical considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission,
redundancy, etc.) have been completely observed by the authors.

Acknowledgements

This study was a part of the National Research Program "Healthy Foods for a Strong Bioeconomy and Quality of Life", with the participation of Medical Faculty, Trakia University, Stara Zagora.

Conflict of interest

The authors declare that there is no conflict of interest.

References


41. Health Policy Institute (2019). Obesity Among Older Americans. hpi.georgetown.edu/obesity2/.


